Title	確率伝搬法による近似と情報幾何学による解析(情報物理学の数学的構造)
Author(s)	Ikeda, Shiro
Citation	数理解析研究所講究録 (2007), 1532: 1-10
Issue Date	2007-02
URL	http://hdl.handle.net/2433/58962
Right	
Type	Departmental Bulletin Paper
Textversion	publisher

数理解析研究所講究録
1532巻 2007年 1-10

確率伝搬法による近似と情報幾何学による解析

統計数理研究所　池田思朗 (Shiro Ikeda)
The Institute of Statistical Mathematics

1 Introduction

We have developed an information geometrical framework of the Belief propagation (BP) algorithm in [1, 2]. One important aspect of the information geometrical framework is that we can view different methods developed in different fields from a unified viewpoint. In this draft, we first summarize the information geometrical framework, and show that the adaptive TAP[3] method is also expressed in the same framework. This will allow us to apply well-developed tools and results of the loopy BP to the adaptive TAP.

2 Information geometrical framework of BP

First, we summarize the information geometrical view of BP.

2.1 Problem

The distribution we would like to focus on is given as follows

$$q(\boldsymbol{x}) = \exp[c_0(\boldsymbol{x}) + c_1(\boldsymbol{x}) + \cdots + c_L(\boldsymbol{x}) - \psi_q].$$

For a while, we restrict ourselves to the case where $\boldsymbol{x} \in \{-1, +1\}^N$. In the case of the Boltzmann machine,

$$c_0(\boldsymbol{x}) = \boldsymbol{h} \cdot \boldsymbol{x}, \quad c_r(\boldsymbol{x}) = J_{ij} x_i x_j.$$

For a large N, we cannot compute ψ_q, which is the log-partition function.

The goal of the BP algorithm is to infer $\boldsymbol{\eta} = E_q[\boldsymbol{x}]$, or equivalently compute $\prod_i q(x_i)$. The direct computation is not tractable when N is large and the graph is cyclic.

2.2 Models, manifolds, and equilibrium

Let us define the set of all distributions S.

$$S = \Big\{ p(\boldsymbol{x}) \mid \sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1, p(\boldsymbol{x}) > 0 \Big\}.$$

A submanifold $M_0 \subset S$ is defined as

$$M_0 = \Big\{ p_0(\boldsymbol{x}; \boldsymbol{\theta}) = \exp[\boldsymbol{h} \cdot \boldsymbol{x} + \boldsymbol{\theta} \cdot \boldsymbol{x} - \psi_0(\boldsymbol{\theta})] \mid \boldsymbol{h}, \boldsymbol{\theta} \in \Re^N \Big\},$$

where $\cdot$ shows the inner product. This is a set of factorizable distributions. Each component is independent for the distributions of M_0, and its natural parameter is $\boldsymbol{\theta}$. Conversely, every factorizable distribution is included in M_0. Therefore, the problem to compute $\prod_i q(x_i)$ is equivalent to compute the natural parameter $\boldsymbol{\theta}$ which satisfies $p_0(\boldsymbol{x}; \boldsymbol{\theta}) = \prod_i q(x_i)$.

Let us define the m–projection of distribution $r(\boldsymbol{x})$ to M_0 as

$$\boldsymbol{\theta} = \pi_{M_0} \circ r(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{\theta}} D[r(\boldsymbol{x}); p_0(\boldsymbol{x}; \boldsymbol{\theta})], \quad \Pi_{M_0} \circ r(\boldsymbol{x}) = \operatorname*{argmin}_{p(\boldsymbol{x}) \in M_0} D[r(\boldsymbol{x}); p_0(\boldsymbol{x}; \boldsymbol{\theta})],$$

where $D[\cdot;\cdot]$ is the KL-divergence, and we have the following relation

$$p_0(\boldsymbol{x}; \boldsymbol{\theta}^*) = \Pi_{M_0} \circ q(\boldsymbol{x}) = \prod_{i=1}^{N} q(x_i), \quad \text{where} \quad \boldsymbol{\theta}^* = \pi_{M_0} \circ q(\boldsymbol{x}).$$

Now, let us define $p_r(\boldsymbol{x}; \boldsymbol{\zeta}_r)$, $r = 1, \cdots, L$ as,

$$p_r(\boldsymbol{x}; \boldsymbol{\zeta}_r) = \exp[\boldsymbol{h} \cdot \boldsymbol{x} + c_r(\boldsymbol{x}) + \boldsymbol{\zeta}_r \cdot \boldsymbol{x} - \psi_r(\boldsymbol{\zeta}_r)], \quad \boldsymbol{\zeta}_r \in \Re^N, \quad r = 1, \cdots, L.$$

$p_r(\boldsymbol{x}; \boldsymbol{\zeta}_r)$ is an exponential family which includes $c_r(\boldsymbol{x})$.

$$M_r = \Big\{ p_r(\boldsymbol{x}; \boldsymbol{\zeta}_r) \mid \boldsymbol{\zeta}_r \in \Re^N \Big\}, \quad r = 1, \cdots, L.$$

Its natural parameter is $\boldsymbol{\zeta}_r$. We assume the computation of $\pi_{M_0} \circ p_r(\boldsymbol{x}; \boldsymbol{\zeta}_r)$ is tractable for every $\boldsymbol{\zeta}_r \in \Re^N$ and r. With $p_0(\boldsymbol{x}; \boldsymbol{\theta})$ and $p_r(\boldsymbol{x}; \boldsymbol{\zeta}_r)$, $r = 1, \cdots, L$, the BP algorithm is defined as follows,

1. Set $t = 0$, $\boldsymbol{\xi}_r^t = \mathbf{o}$, $\boldsymbol{\zeta}_r^t = \mathbf{o}$, $r = 1, \cdots, L$.
2. Increase t by 1 and update $\boldsymbol{\xi}_r^{t+1}$, $r = 1, \cdots, L$ as follows
$$\boldsymbol{\xi}_r^{t+1} = \pi_{M_0} \circ p_r(\boldsymbol{x}; \boldsymbol{\zeta}_r^t) - \boldsymbol{\zeta}_r^t.$$
3. Update $\boldsymbol{\theta}^{t+1}$ and $\boldsymbol{\zeta}_r^{t+1}$ as follows
$$\boldsymbol{\zeta}_r^{t+1} = \sum_{r' \neq r} \boldsymbol{\xi}_{r'}^{t+1}, \quad \boldsymbol{\theta}^{t+1} = \sum_r \boldsymbol{\xi}_r^{t+1} = \frac{1}{L-1} \sum_r \boldsymbol{\zeta}_r^{t+1}.$$
4. Repeat 2 and 3 until $\{\boldsymbol{\xi}_r^t\}$ converges.

At the equilibrium, we know the following conditions hold [2].

m–condition: $\theta^* = \pi_{M_0} \circ p_r(x; \zeta_r^*)$.

e–condition: $\theta^* = \dfrac{1}{L-1}\sum_{r=1}^{L} \zeta_r^*$, or equivalently $q(x) = \dfrac{1}{Z}\dfrac{\prod_{r=1}^{L} p_r(x;\zeta_r)}{p_0(x;\theta)^{L-1}}$.

We have shown these conditions are satisfied at the equilibrium of BP, TRP, and CCCP in [2].

2.3 Free energy

When the e–condition is satisfied, we have the following relation.

$$q(x) = \frac{1}{Z}\frac{\prod_{r=1}^{L} p_r(x;\zeta_r)}{p_0(x;\theta)^{L-1}}.$$

Moreover, when graph is tree, at the equilibrium of BP, Z becomes 1, that is

$$q(x) = \frac{\prod_{r=1}^{L} p_r(x;\zeta_r^*)}{p_0(x;\theta^*)^{L-1}}.$$

This equation shows that when graph is tree, free energy is expressed as

$$\mathcal{F} = -\psi_q = (L-1)\psi_0(\theta^*) - \sum_{r=1}^{L}\psi_r(\zeta_r^*).$$

This relation does not hold for cyclic graphs, but the right hand side of the equation is called the Bethe free energy. This equation is deeply related to the BP algorithm [2].

$$\mathcal{F}_{Bethe} = (L-1)\psi_0(\theta^*) - \sum_{r=1}^{L}\psi_r(\zeta_r^*).$$

3 Adaptive TAP

The adaptive TAP method is described in detail by Opper and Winther[3]. We follow their results, where notations are slightly modified to make consistency with this memo.

3.1 Problem

We consider the distribution

$$q(x) = \frac{1}{Z}\prod_{r=1}^{N}\rho(x_r)\exp\Big[h\cdot x + \frac{1}{2}x^T J x\Big]. \tag{1}$$

J is a symmetric matrix where diagonal elements are 0. In the paper, ρ takes a lot of kinds of functions, such as Dirac delta function. Actually, by taking $\rho(x_r) = (\delta(x_r-1)+\delta(x_r+1))/2$, $q(x)$ becomes equivalent to the Boltzmann machine.

3.2 Adaptive TAP equations

The aim of the adaptive TAP approach is to infer $E_q[x_r]$ and $E_q[x_r^2]$. Let m_r be the inference of $E_q[x_r]$. A summary of Adaptive TAP equations is given as follow.

$$m_r = \frac{\partial}{\partial h_r} \ln Z_0^{(r)} \tag{2}$$

$$Z_0^{(r)} = \int \rho(x_r) \exp\left[\Big(\sum_{s=1}^{N} J_{rs} m_s - V_r m_r + h_r \Big) x_r + \frac{V_r}{2} x_r^2 \right] dx_r$$

$$\frac{\partial m_r}{\partial h_r} = \frac{\partial^2}{\partial h_r^2} \ln Z_0^{(r)} = [(S-J)^{-1}]_{rr}, \tag{3}$$

$$S = \mathrm{diag}(s_1, \cdots, s_N), \quad \frac{1}{s_r - V_r} = \frac{\partial m_r}{\partial h_r} = [(S-J)^{-1}]_{rr}.$$

Let us define $p_r(x_r)$ as follows,

$$p_r(x_r) = \frac{1}{Z_0^{(r)}} \rho(x_r) \exp\left[\Big(\sum_{s=1}^{N} J_{rs} m_s - V_r m_r + h_r \Big) x_r + \frac{V_r}{2} x_r^2 \right]. \tag{4}$$

From the definition of $Z_0^{(r)}$, $p_r(x_r)$ is a density function of x_r. Now, we can rewrite the adaptive TAP equations (2) and (3) as follows,

$$m_r = \int x_r p_r(x_r) dx_r \tag{5}$$

$$\left[(S-J)^{-1}\right]_{rr} = \int (x_r - m_r)^2 p_r(x_r) dx_r = \int x_r^2 p_r(x_r) dx_r - m_r^2. \tag{6}$$

3.3 Free energy

The free energy, which is the inference of $-\ln Z$ is given as follows at the solutions of adaptive TAP equations.

$$\Phi_1 = \Phi_0 - \frac{1}{2} \boldsymbol{m}^T J \boldsymbol{m} + \frac{1}{2} \ln \det(S-J) - \frac{1}{2} \sum_{r=1}^{N} V_r \chi_{rr} + \frac{1}{2} \sum_{r=1}^{N} \ln \chi_{rr}$$

$$\Phi_0 = -\sum_{r=1}^{N} \ln Z_0^{(r)} + \boldsymbol{m}^T J \boldsymbol{m} + \frac{1}{2} \sum_{r=1}^{N} V_r M_r,$$

$$\chi_{rr} = M_r - m_r^2 = [(S-J)^{-1}]_{rr}.$$

where M_r is the inference of $E_q[x_r^2]$. We can simplify the above free energy as

$$\Phi_1 = -\sum_{r=1}^{N} \ln Z_0^{(r)} + \frac{1}{2} \ln \det(S-J) + \frac{1}{2} \sum_{r=1}^{N} \ln[(S-J)^{-1}]_{rr} + \frac{1}{2} \boldsymbol{m}^T J \boldsymbol{m} - \frac{1}{2} \sum_{r=1}^{N} V_r m_r^2. \tag{7}$$

4 From BP to adaptive TAP

We show that the information geometrical framework of the BP algorithm explains the adaptive TAP equations and free energy.

4.1 Problem

We consider the case where $\rho(x_r)$ is strictly positive for $x_r \in \Re$. This is not true when ρ is the Dirac delta function. Naively, we can treat such a problem by putting

$$\rho(x_r) = \frac{1}{2\sqrt{2\pi\sigma^2}} \left(\exp\left[-\frac{(x_r - 1)^2}{2\sigma^2} \right] + \exp\left[-\frac{(x_r + 1)^2}{2\sigma^2} \right] \right),$$

and bringing $\sigma^2 \to 0$. But we put this problem beside and assume $\rho(x_r) > 0$. Equation (1) can be rewritten as

$$\begin{aligned} q(\boldsymbol{x}) &= \exp[c_0(\boldsymbol{x}) + c_1(\boldsymbol{x}) + \cdots + c_N(\boldsymbol{x}) - \psi_q], \\ c_0(\boldsymbol{x}) &= \frac{1}{2}\boldsymbol{x}^T J \boldsymbol{x}, \quad c_r(\boldsymbol{x}) = c_r(x_r) = \ln \rho(x_r) + h_r x_r, \quad \psi_q = \ln Z. \end{aligned} \tag{8}$$

Next, we define p_0 as the distribution whose sufficient statistics are x_r, x_r^2, $r = 1, \cdots, N$. Natural choice is the normal distribution, which is defined as

$$\begin{aligned} p_0(\boldsymbol{x}; \boldsymbol{\mu}, S) &= \exp\left[c_0(\boldsymbol{x}) + \boldsymbol{\mu} \cdot \boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T S \boldsymbol{x} - \psi_0(\boldsymbol{\mu}, S) \right], \\ S &= \mathrm{diag}(s_1, \cdots, s_N), \\ \psi_0(\boldsymbol{\mu}, S) &= \frac{N}{2} \ln 2\pi - \ln \det(S - J) + \frac{1}{2}\boldsymbol{\mu}^T (S - J)^{-1} \boldsymbol{\mu}. \end{aligned} \tag{9}$$

From the definition $p_0(\boldsymbol{x}; \boldsymbol{\mu}, S) \sim \mathcal{N}((S - J)^{-1}\boldsymbol{\mu}, (S - J)^{-1})$, where $\mathcal{N}$ shows the density function of a normal distribution. We set M_0 as

$$M_0 = \{ p_0(\boldsymbol{x}; \boldsymbol{\mu}, S) \mid \boldsymbol{\mu} \in \Re^N, S = \mathrm{diag}(s_1, \cdots, s_N), s_i > 0 \}$$

Now, our ultimate goal is to obtain the m–projection of $q(\boldsymbol{x})$ defined as (8) to M_0, which then provides the exact mean and variance of x_r.

4.2 e–condition and m–condition

We define $p_r(\boldsymbol{x}; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r})$ as follows

$$\begin{aligned} p_r(\boldsymbol{x}; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r}) &= \exp\left[c_0(\boldsymbol{x}) + c_r(x_r) + \boldsymbol{\mu}_{\backslash r} \cdot \boldsymbol{x}_{\backslash r} - \frac{1}{2}\boldsymbol{x}_{\backslash r}{}^T S_{\backslash r} \boldsymbol{x}_{\backslash r} - \psi_r(\boldsymbol{\mu}_{\backslash r}, S_{\backslash r}) \right], \\ S_{\backslash r} &= \mathrm{diag}(s_1, \cdots, s_{r-1}, s_{r+1}, \cdots, s_N) \in \Re^{(N-1)\times(N-1)}, \\ \boldsymbol{\mu}_{\backslash r} &= (\mu_1, \cdots, \mu_{r-1}, \mu_{r+1}, \cdots, \mu_N)^T \in \Re^{N-1}, \\ \boldsymbol{x}_{\backslash r} &= (x_1, \cdots, x_{r-1}, x_{r+1}, \cdots, x_N)^T \in \Re^{N-1}. \end{aligned}$$

We define M_r as

$$M_r = \{p_r(\boldsymbol{x}; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r})\}.$$

The difference between p_0 and p_r is that,

$$\text{"}p_r \textit{ does not include } \mu_r x_r \textit{ nor } -s_r x_r^2/2, \textit{ but includes } c_r(x_r).\text{"}$$

This is similar to the information geometrical framework of the BP algorithm.

e–condition

It is easy to show the *e*–condition holds, that is

$$q(\boldsymbol{x}) = \frac{1}{Z} \frac{\prod_{r=1}^N p_r(\boldsymbol{x}; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r})}{p_0(\boldsymbol{x}; \boldsymbol{\mu}, S)^{N-1}}. \tag{10}$$

The numerator of the right hand side is

$$\prod_{r=1}^N p_r(\boldsymbol{x}; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r}) \propto \exp\left[N c_0(\boldsymbol{x}) + \sum_{r=1}^N c_r(x_r) + (N-1) \sum_{r=1}^N \mu_r x_r - \frac{(N-1)}{2} \sum_{r=1}^N s_r x_r^2 \right],$$

and its denominator is

$$p_0(\boldsymbol{x}; \boldsymbol{\mu}, S)^{N-1} \propto \exp\left[(N-1) c_0(\boldsymbol{x}) + (N-1) \sum_{r=1}^N \mu_r x_r - \frac{(N-1)}{2} \sum_{r=1}^N s_r x_r^2 \right],$$

which proves (10).

m–condition

Next, we show that the *m*–condition corresponds to the adaptive TAP equations. When the *m*–condition holds,

$$p_0(\boldsymbol{x}; \boldsymbol{\mu}, S) = \Pi_{M_0} \circ p_r(\boldsymbol{x}; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r}) = \underset{p(\boldsymbol{x}) \in M_0}{\operatorname{argmin}} D[p_r(\boldsymbol{x}; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r}); p(\boldsymbol{x})], \quad r = 1, \cdots, N.$$

The sufficient statistics of p_0 is x_r, x_r^2, $r = 1, \cdots, N$. Therefore the *m*–condition is equivalent to

$$\boldsymbol{m} = K^{-1}\boldsymbol{\mu} = \int \boldsymbol{x} p_r(\boldsymbol{x}; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r}) d\boldsymbol{x},$$

$$\frac{1}{[K^{-1}]_{ss}} = \int x_s^2 p_r(\boldsymbol{x}; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r}) d\boldsymbol{x} - m_s^2, \quad r, s = 1, \cdots, N.$$

where, $K = S - J$

Let us set $\boldsymbol{J}_r$, $K_{\backslash r}$, and $J_{\backslash r}$ as follows

$$J_{\backslash r} = \begin{pmatrix} 0 & \cdots & J_{1(r-1)} & J_{1(r+1)} & \cdots & J_{1N} \\ \vdots & \ddots & & & & \vdots \\ J_{(r-1)1} & & & & & J_{(r-1)N} \\ J_{(r+1)1} & & & & & J_{(r+1)N} \\ \vdots & & & & \ddots & \vdots \\ J_{N1} & \cdots & J_{N(r-1)} & J_{N(r+1)} & \cdots & 0 \end{pmatrix} \in \Re^{(N-1)\times(N-1)}$$

$$K_{\backslash r} = (S_{\backslash r} - J_{\backslash r}) \in \Re^{(N-1)\times(N-1)}$$

$$\boldsymbol{J}_r = (J_{r1}, \cdots, J_{r(r-1)}, J_{r(r+1)}, \cdots, J_{rN})^T \in \Re^{N-1}$$

We can rewrite $p_r(\boldsymbol{x}; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r})$ as follows,

$$\begin{aligned} p_r(\boldsymbol{x}; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r}) &= p_r(x_r; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r}) p_r(\boldsymbol{x}_{\backslash r} | x_r; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r}) \\ &= p_r(x_r; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r}) \mathcal{N}({K_{\backslash r}}^{-1}(\boldsymbol{\mu}_{\backslash r} + \boldsymbol{J}_r x_r), {K_{\backslash r}}^{-1}) \end{aligned} \tag{11}$$

where,

$$p_r(x_r; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r}) = \frac{1}{Z_r} \exp\Big[c_r(x_r) + \frac{1}{2}(\boldsymbol{J}_r{}^T {K_{\backslash r}}^{-1} \boldsymbol{J}_r) x_r^2 + (\boldsymbol{J}_r{}^T {K_{\backslash r}}^{-1} \boldsymbol{\mu}_{\backslash r}) x_r\Big]. \tag{12}$$

We can also rewrite $p_0(\boldsymbol{x}; \boldsymbol{\mu}, S)$ as follows,

$$\begin{aligned} p_0(\boldsymbol{x}; \boldsymbol{\mu}, S) &= p_0(x_r; \boldsymbol{\mu}, S) p_0(\boldsymbol{x}_{\backslash r} | x_r; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r}) \\ &= p_0(x_r; \boldsymbol{\mu}, S) \mathcal{N}({K_{\backslash r}}^{-1}(\boldsymbol{\mu}_{\backslash r} + \boldsymbol{J}_r x_r), {K_{\backslash r}}^{-1}) \end{aligned} \tag{13}$$

where,

$$\begin{aligned} p_0(x_r; \boldsymbol{\mu}, S) &= \frac{1}{Z_r} \exp\Big[-\frac{1}{2} s_r x_r^2 + \mu_r x_r + \frac{1}{2}(\boldsymbol{J}_r{}^T {K_{\backslash r}}^{-1} \boldsymbol{J}_r) x_r^2 + (\boldsymbol{J}_r{}^T {K_{\backslash r}}^{-1} \boldsymbol{\mu}_{\backslash r}) x_r\Big] \\ &= \mathcal{N}\left(\frac{1}{s_r - \boldsymbol{J}_r{}^T {K_{\backslash r}}^{-1} \boldsymbol{J}_r}(\mu_r + \boldsymbol{J}_r{}^T {K_{\backslash r}}^{-1} \boldsymbol{\mu}_{\backslash r}), \frac{1}{s_r - \boldsymbol{J}_r{}^T {K_{\backslash r}}^{-1} \boldsymbol{J}_r}\right) \end{aligned}$$

And we have the following proposition

proposition 1. *The m-condition is satisfied, if and only if the following equations hold.*

$$m_r = \int x_r p_r(x_r; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r}) dx_r \tag{14}$$

$$[(S - J)^{-1}]_{rr} = \frac{1}{s_r - \boldsymbol{J}_r{}^T {K_{\backslash r}}^{-1} \boldsymbol{J}_r} = \int x_r^2 p_r(x_r; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r}) dx_r - m_r^2 \tag{15}$$

$$r = 1, \cdots, N, \quad \text{where} \quad \boldsymbol{m} = (S - J)^{-1} \boldsymbol{\mu}.$$

Now we move to the adaptive TAP equations

lemma 1. *Equations* (14) *and* (15) *are equivalent to* (5) *and* (6).

Proof. What we have to show is that $p_r(x_r)$ in (4) is equivalent to $p_r(x_r; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r})$ in (12). We show both of them again.

$$p_r(x_r) = \frac{1}{Z_0^{(r)}}\rho(x_r)\exp\left[\left(\sum_{j=1}^{N} J_{rj}m_j - V_r m_r + h_r\right)x_i + \frac{V_r}{2}x_r^2\right] \tag{16}$$

$$\begin{aligned} p_r(x_r; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r}) &= \frac{1}{Z_r}\exp\left[c_r(x_r) + \frac{1}{2}(\boldsymbol{J}_r{}^T K_{\backslash r}{}^{-1}\boldsymbol{J}_r)x_r^2 + (\boldsymbol{J}_r{}^T K_{\backslash r}{}^{-1}\boldsymbol{\mu}_{\backslash r})x_r\right] \\ &= \frac{1}{Z_r}\rho(x_r)\exp\left[(\boldsymbol{J}_r{}^T K_{\backslash r}{}^{-1}\boldsymbol{\mu}_{\backslash r} + h_r)x_r + \frac{1}{2}(\boldsymbol{J}_r{}^T K_{\backslash r}{}^{-1}\boldsymbol{J}_r)x_r^2\right] \end{aligned} \tag{17}$$

$$[(S-J)^{-1}]_{rr} = \frac{1}{s_r - V_r}, \quad \boldsymbol{m}_{\backslash r} = (m_1, \cdots, m_{r-1}, m_{r+1}, \cdots, m_N)^T \in \Re^{N-1}. \tag{18}$$

where we used $c_r(x_r) = \ln\rho(x_r) + h_r x_r$. And by comparing (16) and (17), what we have to prove is

$$\boldsymbol{J}_r{}^T K_{\backslash r}{}^{-1}\boldsymbol{\mu}_{\backslash r} = \sum_{s=1}^{N} J_{rs}m_s - V_r m_r = \boldsymbol{J}_r{}^T \boldsymbol{m}_{\backslash r} - V_r m_r \tag{19}$$

$$\boldsymbol{J}_r{}^T K_{\backslash r}{}^{-1}\boldsymbol{J}_r = V_r. \tag{20}$$

From (15) and (18), $V_r = \boldsymbol{J}_r{}^T K_{\backslash r}{}^{-1}\boldsymbol{J}_r$ is shown and (20) is proved, and from the relation $\boldsymbol{m} = K^{-1}\boldsymbol{\mu}$ or $\boldsymbol{\mu} = K\boldsymbol{m}$, we have

$$\left(\begin{array}{c}\mu_r \\ \boldsymbol{\mu}_{\backslash r}\end{array}\right) = \left(\begin{array}{c|c} s_r & -\boldsymbol{J}_r{}^T \\ \hline -\boldsymbol{J}_r & K_{\backslash r}\end{array}\right)\left(\begin{array}{c} m_r \\ \boldsymbol{m}_{\backslash r}\end{array}\right)$$

$$\begin{aligned} \mu_r &= s_r m_r - \boldsymbol{J}_r \boldsymbol{m}_{\backslash r} \\ \boldsymbol{\mu}_{\backslash r} &= -\boldsymbol{J}_r m_r + K_{\backslash r}\boldsymbol{m}_{\backslash r}. \end{aligned}$$

Which proves (19), that is,

$$\begin{aligned} \boldsymbol{J}_r{}^T K_{\backslash r}{}^{-1}\boldsymbol{\mu}_{\backslash r} &= -\boldsymbol{J}_r{}^T K_{\backslash r}{}^{-1}\boldsymbol{J}_r m_r + \boldsymbol{J}_r{}^T \boldsymbol{m}_{\backslash r} \\ &= \boldsymbol{J}_r{}^T \boldsymbol{m}_{\backslash r} - V_r m_r. \end{aligned}$$

Thus, the m–condition derives the adaptive TAP equations. □

4.3 Free energy

Since we have seen that the framework of the adaptive TAP method is very similar to that of the BP algorithm, we can define the free energy as the Bethe free energy, that is

$$\mathcal{F} = (N-1)\psi_0(\boldsymbol{\mu}, S) - \sum_{r=1}^{N}\psi_r(\boldsymbol{\mu}_{\backslash r}, S_{\backslash r}).$$

We now show that this free energy is equivalent to the adaptive TAP free energy in (7). Above equation can be rewritten as

$$\mathcal{F} = \sum_{r=1}^{N}(\psi_0(\boldsymbol{\mu}, S) - \psi_r(\boldsymbol{\mu}_{\backslash r}, S_{\backslash r})) - \psi_0(\boldsymbol{\mu}, S).$$

From (9),

$$\begin{aligned}\psi_0(\boldsymbol{\mu}, S) &= \frac{N}{2}\ln 2\pi - \ln\det(S-J) + \frac{1}{2}\boldsymbol{\mu}^T(S-J)^{-1}\boldsymbol{\mu} \\ &= \frac{N}{2}\ln 2\pi - \ln\det(S-J) + \frac{1}{2}\boldsymbol{m}^T(S-J)\boldsymbol{m}\end{aligned}$$

And by comparing (11) and (13), we have

$$\begin{aligned}\psi_0(\boldsymbol{\mu}, S) - \psi_r(\boldsymbol{\mu}_{\backslash r}, S_{\backslash r}) &= -\ln Z_r + \frac{1}{2}\ln 2\pi - \frac{1}{2}\ln(s_r - \boldsymbol{J}_r{}^T K_{\backslash r}{}^{-1}\boldsymbol{J}_r) \\ &\quad + \frac{1}{s_r - \boldsymbol{J}_r{}^T K_{\backslash r}{}^{-1}\boldsymbol{J}_r}(\mu_r + \boldsymbol{J}_r{}^T K_{\backslash r}{}^{-1}\boldsymbol{\mu}_{\backslash r})^2 \\ &= -\ln Z_r + \frac{1}{2}\ln 2\pi + \frac{1}{2}\ln[(S-J)^{-1}]_{rr} + \frac{1}{2}\frac{1}{[(S-J)^{-1}]_{rr}}m_r^2\end{aligned}$$

The free energy is rewritten as

$$\begin{aligned}\mathcal{F} &= -\sum_{r=1}^N \ln Z_r + \ln\det(S-J) - \frac{1}{2}\boldsymbol{m}^T(S-J)\boldsymbol{m} + \frac{1}{2}\sum_{r=1}^N \ln[(S-J)^{-1}]_{rr} + \frac{1}{2}\sum_{r=1}^N \frac{1}{[(S-J)^{-1}]_{rr}}m_r^2 \\ &= -\sum_{r=1}^N \ln Z_r + \ln\det(S-J) + \frac{1}{2}\sum_{r=1}^N \ln[(S-J)^{-1}]_{rr} + \frac{1}{2}\boldsymbol{m}^T J\boldsymbol{m} - \frac{1}{2}\sum_{r=1}^N V_r m_r^2.\end{aligned}$$

This is equivalent to (7).

4.4 BP like algorithm for adaptive TAP

It is written in [3], that it is not easy to compute the solution of adaptive TAP equations. But since the framework of adaptive TAP method is similar to that of the BP, we can derive BP or related algorithms such as CCCP to solve the adaptive TAP equations. Here we show one example, which is the BP like algorithm to solve the adaptive TAP equations.

1. Set $t = 0$ and $\boldsymbol{\mu}^{(t)} = \mathbf{o}$, $s_r^{(t)} = c$, ($c > 0$, for example, $c = 1$) $r = 1, \cdots, L$.
2. Increase t by 1 and calculate $m_r^{(t)}$ and ${\sigma_r^{(t)}}^2$ $r = 1, \cdots, L$ as follows

$$\begin{aligned}m_r^{(t)} &= \int x_r p_r(x_r; {\boldsymbol{\mu}_{\backslash r}}^{(t)}, {S_{\backslash r}}^{(t)})dx_r \\ {\sigma_r^{(t)}}^2 &= \int (x_r - m_r^{(t)})^2 p_r(x_r; {\boldsymbol{\mu}_{\backslash r}}^{(t)}, {S_{\backslash r}}^{(t)})dx_r\end{aligned}$$

where $p_r(x_r; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r}) = \frac{1}{Z_r}\exp\Big[c_r(x_r) + \frac{1}{2}(\boldsymbol{J}_r{}^T K_{\backslash r}{}^{-1}\boldsymbol{J}_r)x_r^2 + (\boldsymbol{J}_r{}^T K_{\backslash r}{}^{-1}\boldsymbol{\mu}_{\backslash r})x_r\Big]$.

3. Update $\mu_r^{(t+1)}$ and $s_r^{(t+1)}$ $r = 1, \cdots, L$ as follows

$$\begin{aligned}s_r^{(t+1)} &= \frac{1}{{\sigma_r^{(t)}}^2} + {\boldsymbol{J}_r}^{(t)T} {K_{\backslash r}}^{(t)-1} {\boldsymbol{J}_r}^{(t)} \\ \mu_r^{(t+1)} &= \frac{m_r^{(t)}}{{\sigma_r^{(t)}}^2} + {\boldsymbol{J}_r}^{(t)T} {K_{\backslash r}}^{(t)-1} {\boldsymbol{\mu}_{\backslash r}}^{(t)}.\end{aligned}$$

4. Repeat 2 and 3 until convergence.

(We have no idea if this algorithm works or not.)

5 Summary

We have shown the information geometrical framework of the belief propagation algorithm and that the adaptive TAP equations are derived in the similar manner as the equilibrium conditions of the BP algorithm. Here we defined $q(\boldsymbol{x})$, $p_0(\boldsymbol{x};\boldsymbol{\mu},S)$, and $p_r(\boldsymbol{x};\boldsymbol{\mu}_{\backslash r},S_{\backslash r})$ in a different way from original BP, but the information geometrical view is almost the same for both cases. We have also shown that the free energy, which corresponds to the Bethe free energy in the case of BP, is the adaptive TAP free energy.

In the case of BP, the solution becomes exact when the graph is tree. In this problem, we do not have the same result, but adaptive TAP becomes exact when the target distribution $q(\boldsymbol{x})$ is a normal distribution. The information geometrical framework immediately shows this since when $q(\boldsymbol{x})$ is a normal distribution, $q(\boldsymbol{x}), p_0(\boldsymbol{x};\boldsymbol{\mu},S), p_r(\boldsymbol{x};\boldsymbol{\mu}_{\backslash r},S_{\backslash r}) \in M_0$, and every distribution becomes equivalent.

This memo shows that the information geometrical frameworks of BP and adaptive TAP are equivalent. From this fact, there is a possibility that we can reuse a lot of results developed for BP and inference problem of loopy graphs. One important direction is the algorithm. We can derive the BP and CCCP algorithms to compute the equilibrium of adaptive TAP equations.

References

[1] Shiro Ikeda, Toshiyuki Tanaka, and Shun-ichi Amari. Information geometry of turbo and low-density parity-check codes. *IEEE Transactions on Information Theory*, 50(6):1097–1114, June 2004.

[2] Shiro Ikeda, Toshiyuki Tanaka, and Shun-ichi Amari. Stochastic reasoning, free energy, and information geometry. *Neural Computation*, 16(9):1779–1810, September 2004.

[3] Manfred Opper and Ole Winther. Adaptive and self-averaging Thouless-Anderson-Palmer mean field theory for probabilistic modeling. *Physical Review E*, 64(5):056131, 2001.